Armero, C., Cabras, S., Castellanos, M. E., Perra, S., Quirós, A., Oruezábal, M. J. & Sánchez-Rubio, J. (2016). Bayesian analysis of a disability model for lung cancer survival. *Statistical Methods in Medical Research, 25*(1), pp. 336-351.

DOI: 10.1177%2F0962280212452803

Bayesian analysis of a disability model for lung cancer survival

C Armero,[1] S Cabras,[2,3] ME Castellanos,[4] S Perra,[2] A Quirós,[5] MJ Oruezábal[6,7] and J Sánchez-Rubio[8]

Abstract

Bayesian reasoning, survival analysis and multi-state models are used to assess survival times for Stage IV non-small-cell lung cancer patients and the evolution of the disease over time. Bayesian estimation is done using minimum informative priors for the Weibull regression survival model, leading to an automatic inferential procedure. Markov chain Monte Carlo methods have been used for approximating posterior distributions and the Bayesian information criterion has been considered for covariate selection. In particular, the posterior distribution of the transition probabilities, resulting from the multi-state model, constitutes a very interesting tool which could be useful to help oncologists and patients make efficient and effective decisions.

Keywords
Accelerated failure time models, Bayesian information criterion, minimum informative prior, multi-state models, Weibull distribution

1 Introduction

Lung cancer is the leading cause of cancer death in developed countries; for instance, it is the second incident malignant neoplasia in Spain. Despite all new advances in its treatment, 5-year absolute survival rate is currently only 10.2%,[1] and consequently, new therapeutic strategies based on suitable prognostic factors constitute an important piece of research. From a clinical perspective, it would be interesting to personalize the most adequate therapeutic option for each patient.

Although the usual procedures to select the most appropriate treatment for a patient suffering cancer are based on clinical trials, the National Cancer Institute and the National Institute for

Departament d'Estadística i Investigació Operativa, Universitat de València, Spain
[2]Dipartimento di Matematica e Informatica, Universitá degli Studi di Cagliari, Italy
[3]Departamento de Estadística, Universidad Carlos III de Madrid, Spain
[4]Departamento de Estadística e Investigación Operativa, Universidad Rey Juan Carlos, Spain
[5]Departamento de Teoría de la Señal y Comunicaciones, Universidad Rey Juan Carlos, Spain
[6]Unidad Onco-Hematológica, Hospital Universitario Infanta Cristina de Madrid, Spain
[7]Servicio Oncología Médica, Hospital Rey Juan Carlos, Spain
[8]Servicio de Farmacia, Hospital Universitario Infanta Cristina de Madrid, Spain

Corresponding author:
A Quirós, Universidad Rey Juan Carlos, Departamental III 012 Camino del Molino, s/n. 28943 Fuenlabrada (Madrid), Spain.
Email: alicia.quiros@urjc.es

Health and Clinical Excellence agree in promoting complementary designs and methodologies, such as observational studies, to help improve therapy decision making. In particular, understanding the role and significance of prognostic factors in cancer would become a tool to gain knowledge about the different features of this disease: survival times, progression of the disease, response to treatment, complications, etc., where the first two are the characteristics of interest in this article.

Within this framework, we focus on data corresponding to Stage IV non-small-cell lung cancer (*NSCLC*), which is the most prevalent type of lung cancer.[2] The severity of the disease depends on different factors such as tumour size, involvement of lymph nodes and metastasis to other parts of the body. This type of cancer is graded in four stages, denoting the presence of metastatic disease as Stage IV. It is not curable, but treatable. At this stage and after therapy, the medical protocol establishes a periodical monitoring of the patients, but the seriousness of the disease causes, in practice, most of them have a quasi-continuous follow-up. In this stage, the time to tumour progression, called progression-free survival, indicates when the disease recurs. As cancers typically grow before they cause death, there are markers that provide readouts of tumour growth often considerably before the patients die of tumour. With the newer treatments, lack of progression may be associated with a good improvement in outcome. Therefore, a positive answer would be no tumour progression, regardless of the degree of reduction of tumour size and a negative outcome is death or tumour progression. Death without progression, also called non-cancer death, is mainly caused by respiratory and cardiovascular failures together with damages caused by cancer treatment.[3]

In this article, survival times and disease progression for Stage IV *NSCLC* patients in conjunction with some risk factors have been examined using multi-state models and Bayesian reasoning. Multi-state models are a class of stochastic processes which model the probability of visiting a certain set of discrete states in continuous time (Andersen et al.[4] is the seminal reference). They are the natural stochastic models for describing the evolution over time of discrete systems, thus being very attractive for dealing with longitudinal failure time data. Medicine and public health is a breeding ground for them with many relevant statistical methodological contributions and applications,[5–11] some of which are within a Bayesian setting.[12,13] Some updated reviews can be found in Hougaard[14] and Putter et al.[15]

Multi-state models allow for different structures depending on the number and relationships between the states of the processes. Here, we concentrate on the disability model, or illness–death model, a specific class of multi-state models, which is relevant in irreversible diseases where, as for Stage IV *NSCLC* patients, a significant illness' progression increases the risk of death.[14] In particular, we assume that the survival time for a patient who has experienced a tumour progression depends on the time spent in the progression state, thus working in the framework of the homogeneous semi-Markov models.

Survival analysis and multi-state models are closely connected: transition intensities between states in stochastic processes correspond to hazard rate functions for times between transitions in survival analysis.[8] Weibull accelerated failure time models[16] are considered for dealing with the different transition times and Bayesian inference for the Weibull distribution parameters is carried out using minimum informative priors.[17] Markov chain Monte Carlo (*MCMC*) methods have been used for approximating posterior distributions and the Bayesian information criterion (*BIC*) version for censored data proposed in Volinsky and Raftery[18] has been considered for covariate selection. The posterior uncertainty about the parameters of the model can be propagated to approximate not only the posterior distribution for the relevant hazard rate functions and associated quantities but also the posterior distribution for all different transition probabilities, a very interesting tool which could be useful to help oncologists and patients make efficient and effective decisions.

The structure of the article is as follows. Section 2 contains a description of the available data. Section 3 introduces the disability model. Sections 4 and 5 are both methodological and applied.

Section 4 examines the main elements of the Bayesian inference process and covariate selection for survival times and discusses final model assessment for the data. Posterior distributions for transition probabilities are defined in Section 5. In addition, in this section, their performance and characteristics for different generic type of patients are shown. Conclusions and further remarks are in Section 6.

2 Stage IV NSCLC data

Stage IV is an important phase of *NSCLC* because at the time of diagnosis of the disease, between 40% and 50% of patients present incurable metastatic disease and are diagnosed with Stage IV.[19] The variability of results observed in different patients in previous studies[20,21] on the survival times suggests that advanced *NSCLC* is a heterogeneous disease related to a wide spectrum of clinical features. Therefore, further research on identifying suitable prognostic factors could provide valuable information to define *NSCLC* patient subgroups with a similar survival potential.

The data considered in this article are provided by the Infanta Cristina Hospital of Madrid (Spain) and come from a longitudinal study started on 1 January 2008 and still active. Only for the purpose of this study it is considered as finished on 31 December 2010. Time-on-study is assumed as the time scale[22,23] and diagnosis time in Stage IV is fixed as the starting time for each patient.

A total of 35 subjects have been followed and, for each of them, survival times related to progression and overall survival have been collected. All patients received a conventional chemotherapy treatment. There are 9 patients who have died without experiencing any progression of the disease and among the 13 who have progressed, 7 have died. Considering that a patient who has not progressed is censored for overall survival before progression, then a total of 26 (74%) patients are censored for overall survival before progression. Similarly, as dead patients are censored for progression-free survival, 22 (63%) are censored for the progression-free survival. As there are 6 living patients out of the 13 who have progressed, thus 46% are censored for the overall survival after progression.

Independent right-censoring is assumed and, consequently, censored patients have been considered as representative of all non-censored patients who have the same values for the explanatory variables. Although possibly a bit more realistic, we have discarded interval censoring because, as we have mentioned before, the severity of the disease and the special characteristics of the current Spanish Public Health System make the patients to be quasi-continuously followed.

We have initially considered 14 baseline prognostic factors or covariates, suggested by the oncologists: age, gender, smoking habit, body mass index (Bmi), baseline state and previous complications with regard to patient information; tumour location, number of affected organs, histological type and baseline analytics with regard to tumour characteristics; and finally, carcinoembryonic antigen (Cea), lactate dehydrogenase (Ldh), anaemia, calcaemia (Cal) and albumin (Alb) for baseline analytical. In order to provide more insight on the available sample, Tables 1 and 2 present summary statistics of such covariates.

3 The disability model

Let $Z(t)$ be the stochastic process describing the state of a Stage IV *NSCLC* patient at time t, where t is time since diagnosis in Stage IV. All living patients that have not yet progressed are considered to be in state 1, state 2 corresponds to the progression of the tumour, and state 3 stands for the death of the patient. The state space is thus $\{1, 2, 3\}$, states 1 and 2 being transient and state 3 absorbing.

Table 1. Number of patients and percentage (within parentheses), of the categorical covariates groups for the 35 patients in the study

	%		%
Gender		Smoking habit	
Female	6 (17)	Smoker or ex-smoker	30 (86)
Male	29 (83)	Non-smoker	5 (14)
Baseline state		Number of affected organs	
0–1	18 (52)	1	27 (77)
2	6 (17)	2	7 (20)
NA	11 (31)	3	1 (3)
Tumour location		Histological type	
Hilar mass	16 (46)	Adenocarcinoma	10 (29)
Peripheral mass	13 (37)	Squamous	6 (17)
Multi-nodular	6 (17)	Undetermined	19 (54)
Number of complications			
None	11 (31)		
One or more	24 (69)		

Table 2. Median and range of the continuous covariates for the 35 patients in the study

	Median	Range
Age	63.0	49–82
Body mass index (Bmi)	24.8	17.3–30.1
Albumin (Alb)	3.5	2.1–4.6
Anaemia	12.8	9.6–16.3
Calcaemia (Cal)	9.6	8.8–10.8
Carcinoembryonic antigen (Cea)	2.9	0.5–8357.4
Lactate dehydrogenase (Ldh)	298.0	147–2744

Transitions between states are determined by changes in the patient's health condition: from state 1 to 2, $1 \to 2$, when the progression of the cancer is observed, $2 \to 3$ ($1 \to 3$) when a patient who has (not) experienced a progression of the tumour dies. Figure 1 schematically represents this structure.

Transition probabilities between states depend on a vector of parameters, $\boldsymbol{\theta}$, and a set of covariates, $\mathbf{x}$. The probability transition $2 \to 3$ also depends on the sojourn time in state 1, T_{12}

$$p_{1j}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) = P(Z(t) = j \,|Z(s) = 1, \mathbf{x}, \boldsymbol{\theta}), s \leq t;\ j = 2, 3$$
$$p_{23}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, t_{12}) = P(Z(t) = 3|Z(s) = 2, \mathbf{x}, \boldsymbol{\theta}, T_{12} = t_{12}),\ t_{12} \leq s \leq t$$

These probabilities can be determined from transition intensities between states that correspond, in the survival analysis arena, to hazard rate functions for times between transitions, T_{ij}

$$h_{1j}(t \mid \mathbf{x}_{1j}, \boldsymbol{\theta}_{1j}) = \lim_{\triangle t \to 0} \frac{P(t \leq T_{1j} < t + \triangle t \mid T_{1j} \geq t, \mathbf{x}_{1j}, \boldsymbol{\theta}_{1j})}{\triangle t},\ j = 2, 3$$

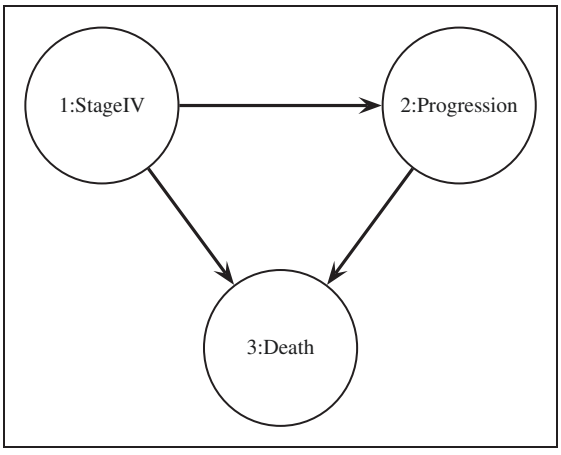

Figure 1. The disability model.

$$h_{23}(t - t_{12} \mid \mathbf{x}_{23}, \boldsymbol{\theta}_{23}, t_{12}) = \lim_{\triangle t \to 0} \frac{P(t - t_{12} \leq T_{23} < t - t_{12} + \triangle t \mid T_{23} \geq t - t_{12}, \mathbf{x}_{23}, \boldsymbol{\theta}_{23}, T_{12} = t_{12})}{\triangle t}$$

where T_{13} is time to death without progression, T_{12} time to progression, and T_{23} time from progression to death. Note that the semi-Markov assumption makes $h_{23}(\cdot)$ to be defined as a function of the sojourn time in state 2 and not of the time since diagnosis.

In the case of the homogeneous semi-Markov disability model, transition probabilities and hazard rate functions are connected as follows[6,24]

$$\begin{aligned}
p_{11}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) &= \exp\left\{-\int_s^t [h_{12}(u \mid \mathbf{x}_{12}, \boldsymbol{\theta}_{12}) + h_{13}(u \mid \mathbf{x}_{13}, \boldsymbol{\theta}_{13})]\mathrm{d}u\right\} \\
p_{22}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, t_{12}) &= \exp\left\{-\int_s^t h_{23}(u - t_{12} \mid \mathbf{x}_{23}, \boldsymbol{\theta}_{23}, t_{12})\mathrm{d}u\right\}, \ t_{12} \leq s \\
p_{12}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) &= \int_s^t p_{11}(s, u \mid \mathbf{x}, \boldsymbol{\theta})\, h_{12}(u \mid \mathbf{x}_{12}, \boldsymbol{\theta}_{12})\, p_{22}(u, t \mid \mathbf{x}, \boldsymbol{\theta}, u)\mathrm{d}u \\
p_{13}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) &= 1 - p_{11}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) - p_{12}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) \\
p_{23}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, t_{12}) &= 1 - p_{22}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, t_{12}) \\
p_{33}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) &= 1
\end{aligned} \tag{1}$$

Note that if we were interested in the marginal distribution regarding T_{12}, $p_{22}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$, we would need to integrate $p_{22}(s, t \mid \mathbf{x}, \theta, t_{12})$ with regard to the conditional density of T_{12} given that $T_{12} \leq s$.

In order to directly express the hazard rate functions in terms of the vector of covariates and parameters, it is necessary to assess the distribution of the different transitions (survival) times.

4 Survival times modelling and covariate selection

In general, the likelihood function for multi-state Markovian or semi-Markovian models can be expressed through the product of the likelihood for the possible transitions.[6,14]

For each individual k, $k=1,\ldots, n$, let

- $y^I_{j,k}$ be the time of entrance into state j since diagnosis of Stage IV.
- $y^O_{j,k}$ ($j=1, 2$) be time of departure from state j (time since diagnosis) or the censored time if the patient remains in state j.
- $c_{ij,k}$ be the indicator of transition $i \to j$, i.e. it is 1 if the $i \to j$ transition is observed for patient k and 0 if it is censored.

In particular, if transition $1 \to 2$ is observed for patient k ($c_{12,k}=1$) at time $t_{12,k}$, then $y^O_{1,k} = y^I_{2,k} = t_{12,k}$, but if transition $1 \to 3$ is observed ($c_{13,k}=1$) at $t_{13,k}$, then $y^O_{1,k} = y^I_{3,k} = t_{13,k}$. In the first case, time from Stage IV diagnosis to death for patient k, $T_{13,k}$, is censored at value $t_{12,k}$ with $c_{13,k}=0$, while in the second case, time from diagnosis to progression for patient k, $T_{12,k}$, is censored at value $t_{13,k}$ and $c_{12,k}=0$. If no transition from state 1 occurs, define $y^O_{1,k} = r_k$, the end of the follow-up for patient k ($c_{12,k}=0$ and $c_{13,k}=0$) and $T_{12,k}$ and $T_{13,k}$ are censored at time r_k. Finally, in case of transition $1 \to 2$ at $t_{12,k}$, $y^O_{2,k}$ is either the time of a $2 \to 3$ transition ($c_{23,k}=1$) or $y^O_{2,k} = r_k$ if censored ($c_{23,k}=0$).

Based on the counting process representation of the likelihood for a multi-state process given in Andersen and Keiding,[6] we derive the full likelihood for the homogeneous semi-Markovian model

$$L(\boldsymbol{\theta} \mid \mathbf{y^I}, \mathbf{y^O}, \mathbf{c}, \mathbf{x}) = \prod_{k=1}^{n} f_{12}(y^O_{1,k})^{c_{12k}} S_{12}(y^O_{1,k})^{1-c_{12k}} f_{13}(y^O_{1,k})^{c_{13k}} S_{13}(y^O_{1,k})^{1-c_{13k}} \times \prod_{k \in P} f_{23}(y^O_{2,k} - y^I_{2,k})^{c_{23k}} S_{23}(y^O_{2,k} - y^I_{2,k})^{1-c_{23k}} \tag{2}$$

where $\mathbf{y}^I$ ($\mathbf{y}^O$) is the vector of all entrance (departure) times into (from) the corresponding states of the patients in the sample, $\mathbf{c}$ all the transition indicators, $\mathcal{P}$ the subset of indexes of patients who have experimented progression, $f_{ij}(\cdot)$ the density for transition times T_{ij} and $S_{ij}(\cdot)$

$$S_{ij}(u) = \exp\left(-\int_0^u h_{ij}(u)\,\mathrm{d}u\right)$$

Note that in order to make the expression of the likelihood in equation (2) clearer, $f_{ij}(\cdot)$ and $S_{ij}(\cdot)$ appear without conditioning on the corresponding parameters and covariates. The arguments of these functions can be expressed in terms of T_{12}, T_{13}, and T_{23}, respectively. More details about the construction of this likelihood can be found in Appendix.

In this article, each survival time, T_{ij}, is modelled through the Weibull distribution, a traditional model for survival data[16] that we examine in the framework of accelerated failure time models. We use this model because of its flexibility in representing different types of risks. In addition, the Weibull distribution is a traditional model widely used in biomedical applications for dealing with data involving survival times.[10,11,25,24]

As the inferential procedure for the Weibull family is the same for each transition time, we introduce it in a general way by omitting subindices for survival times. Accordingly, the logarithm of the corresponding survival time, T, can be expressed in terms of a regression model with standard Gumbel error, W

$$\log(T) = \mu + \mathbf{x}'\beta + \sigma W \tag{3}$$

where β is the vector of coefficients associated to covariates $\mathbf{x}'$ and μ and σ the intercept and the scale parameter, respectively.

Given $\mathbf{x}$ and $(\mu, \beta, \sigma)'$, the distribution for T is a Weibull distribution with parameters $\alpha = 1/\sigma$ and $\lambda(\mathbf{x}) = \exp(-(\mu + \mathbf{x}'\beta)/\sigma)$ and thus, the hazard rate function is

$$h(t \mid \mu, \beta, \sigma, \mathbf{x}) = \frac{1}{\sigma} \exp\left\{-\frac{(\mu + \mathbf{x}'\beta)}{\sigma}\right\} t^{1/\sigma - 1} \tag{4}$$

As $\log(T)$ is distributed according to a location-scale model, a natural default prior for (μ, β, σ) is

$$\pi(\mu, \beta, \sigma) \propto \frac{1}{\sigma} \tag{5}$$

This prior has also been discussed in Evans and Nigm[27] and Albert.[28]

Before continuing with the discussion of the full Bayesian analysis and computing the posterior distribution for all the parameters of the multi-sate model, we are going to discuss covariate selection for each transition time. *BIC* has been used to select the covariates significantly related to each transition time separately. *BIC* was derived by Schwarz[29] as a large sample approximation to twice the logarithm of the Bayes factor (*BF*), which quantifies the evidence for one model against another.[30] In our case, we use the version of the *BIC* proposed in Volinsky and Raftery[18] for censored data because the *BF* is undetermined when improper priors are used. Variable selection in our problem requires a great computational cost as, in the context of regression models with p possible covariates, and when we consider each transition separately, that is, different parameters for each transition, the number of possible models is at least 2^p, here resulting in 16,384 models.

First, we have sorted all the possible models according to their *BIC* value. Then, the relationship between the covariates in the model for each transition and survival has been interpreted, discarding those models lacking medical interpretation. It is important to note that discarded models have a *BIC* not significantly higher than those of the proposed ones. Finally, the selected models for transition times are

$$\begin{aligned}
\log(T_{12}) &= \mu_{12} + \beta_{12,\,\mathrm{Alb}} + \beta_{12,\,\mathrm{Bmi}}\mathrm{Bmi} + \beta_{12,\,\mathrm{Cea}}\mathrm{Cea} \\
&\quad + \beta_{12,\,\mathrm{Ldh1}}\mathrm{Ldh1} + \beta_{12,\,\mathrm{Ldh2}}\mathrm{Ldh2} + \sigma_{12}W_{12} \\
\log(T_{13}) &= \mu_{13} + \beta_{13,\,\mathrm{Cal}}\mathrm{Cal} + \sigma_{13}W_{13} \\
\log(T_{23}) &= \mu_{23} + \beta_{23,\,\mathrm{Cal}}\mathrm{Cal} + \sigma_{23}W_{23}
\end{aligned} \tag{6}$$

where all covariates except for Cal and Alb are discretized in order to avoid the effect of extreme observations, and following medical indications. Bmi and Cea has been dichotomized in two groups and Ldh in three (see Table 3 for a complete definition and description). These variables have been included in the model through dummy variables and, consequently, for the three-categorical variable Ldh, two dummies variables, Ldh1 and Ldh2, have been introduced into the design matrix.

Returning to Bayesian estimation for the full multi-state vector of parameters and assuming a prior distribution for them that considers prior independence among the parameters of the different transitions according to the general form in equation (5), the kernel of the posterior distribution is obtained by multiplying the prior by the full likelihood (2). The posterior distribution has been approximated with *MCMC*, in particular a random walk Metropolis–Hastings[31] for all involved parameters (μ_{ij}, β_{ij}, $\log(\sigma_{ij})$) using a multi-variate normal as proposal distribution.

Table 3. Discretized covariates in model (6)

	Number of patients	%
Bmi		
0: $18 \leq \text{Bmi} \leq 25$	16	46
1: $\text{Bmi} < 18$ or $\text{Bmi} > 25$	19	54
Cea		
0: $\text{Cea} \leq 30$	32	91
1: $\text{Cea} > 30$	3	9
Ldh		
0: $\text{Ldh} \leq 250$	18	51
1: $250 < \text{Ldh} \leq 400$	10	29
2: $\text{Ldh} > 400$	7	20

Table 4. Summaries of the simulated values from the posterior distribution for the parameters of each transition time

	Mean	Median	SD	$q_{0.025}$	$q_{0.975}$
$1 \rightarrow 3$					
μ_{13}	6.19	6.15	0.25	5.79	6.77
$\beta_{13,\text{Cal}}$	−0.42	−0.42	0.15	−0.75	−0.13
$\sigma_{13} = \sigma_{12}$	0.57	0.55	0.12	0.39	0.84
$1 \rightarrow 2$					
μ_{12}	7.28	7.21	0.51	6.46	8.44
$\beta_{12,\text{Alb}}$	0.58	0.58	0.22	0.15	1.02
$\beta_{12,\text{Bmi}}$	−1.17	−1.16	0.39	−1.98	−0.43
$\beta_{12,\text{Cea}}$	−1.39	−1.42	0.53	−2.34	−0.22
$\beta_{12,\text{Ldh1}}$	−0.64	−0.60	0.50	−1.73	0.24
$\beta_{12,\text{Ldh2}}$	−0.94	−0.90	0.48	−2.00	−0.09
$2 \rightarrow 3$					
μ_{23}	6.41	6.17	1.25	4.74	9.68
$\beta_{23,\text{Cal}}$	−2.10	−1.81	1.77	−6.32	0.50
σ_{23}	2.34	2.07	1.07	1.10	5.16

Finally, we have investigated several parameter restrictions in order to obtain more precise results and decrease the uncertainty. We have fitted the selected model, (6), when working separately for each transition assuming the following restrictions in the shape parameter of Weibull: (a) $\sigma_{12} = \sigma_{13} = \sigma_{23}$; (b) $\sigma_{12} = \sigma_{13}$ and a different σ_{23}; (c) $\sigma_{13} = \sigma_{23}$ and a different σ_{12}; (d) $\sigma_{12} = \sigma_{23}$ and a different σ_{13}; and (e) all three σ_{ij} different. We have compared these models using *BIC*, and the selected one has been the model (b).

After a burn-in of 10,000 *MCMC* draws, Table 4 presents summaries of the remaining 50,000 draws from the posterior distribution. It is straightforward that the largest uncertainty corresponds to transition probability $2 \rightarrow 3$ because of the small number of patients that have progressed. In both transitions related with death, $1 \rightarrow 3$ and $2 \rightarrow 3$, the summary measures of Cal are negative, meaning the risk of dying increases with increasing value of Cal. For transition

$1 \to 2$, given Alb and Bmi, covariates Cea and Ldh have a negative effect: as they have been discretized, having Cea > 30 and $250 < \text{Ldh} \leq 400$ or Ldh > 400 values increases the risk of progression. Bmi has a similar effect, with Bmi < 18 or Bmi$>$ 25 increasing the risk of progression. Finally, Alb appears to have a moderate protective effect.

The goodness-of-fit of the Weibull parametric model has been investigated using graphical tools. In particular, we have used a diagnostic plot comparing the logarithm of the Nelson–Aalen estimator of the cumulative hazard with respect to the logarithm of time, indicating that the assumed hazard baseline is compatible with the observed data (see Chapter 12 of Klein and Moeschberger[16]). The semi-Markov assumption, that is, $h_{23}(\cdot)$ does not depend on previous history, has been assessed analysing if the time between progression and death, T_{23}, depends on time until progression.[8]

5 Posterior distribution for transition probabilities

Taking into account the expression of the hazard rate function in equation (4) for Weibull survival times and the fitted model, we can express the transition probabilities between states in equation (1) as

$$\begin{aligned}
p_{11}(s,t \mid \mathbf{x},\boldsymbol{\theta}) &= \exp\left\{-\lambda_{12}(\mathbf{x}_{12})\left(t^{\sigma_{12}^{-1}} - s^{\sigma_{12}^{-1}}\right) - \lambda_{13}(\mathbf{x}_{13})\left(t^{\sigma_{13}^{-1}} - s^{\sigma_{13}^{-1}}\right)\right\} \\
p_{22}(s,t \mid \mathbf{x},\boldsymbol{\theta},t_{12}) &= \exp\left\{-\lambda_{23}(\mathbf{x}_{23})\left((t-t_{12})^{\sigma_{23}^{-1}} - (s-t_{12})^{\sigma_{23}^{-1}}\right)\right\} \\
p_{12}(s,t \mid \mathbf{x},\boldsymbol{\theta}) &= \sigma_{12}^{-1}\lambda_{12}(\mathbf{x}_{12})\exp\left\{\lambda_{12}(\mathbf{x}_{12})s^{\sigma_{12}^{-1}} + \lambda_{13}(\mathbf{x}_{13})s^{\sigma_{13}^{-1}}\right\} \\
&\quad \times \int_s^t \left(u^{\sigma_{12}^{-1}-1} \times \exp\left\{-\lambda_{13}(\mathbf{x}_{13})u^{\sigma_{13}^{-1}} - \lambda_{12}(\mathbf{x}_{12})u^{\sigma_{12}^{-1}} - \lambda_{23}(\mathbf{x}_{23})(t-u)^{\sigma_{23}^{-1}}\right\}\right)\mathrm{d}u \quad (7)
\end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_{12},\ \boldsymbol{\theta}_{13},\ \boldsymbol{\theta}_{23})'$, $\boldsymbol{\theta}_{ij} = (\mu_{ij},\ \sigma_{ij},\ \beta_{ij})'$, $i < j \in \{2,\ 3\}$, $\lambda_{ij}(\mathbf{x}_{ij}) = \exp\left\{-\left(\mu_{ij} + \mathbf{x}_{ij}'\beta_{ij}\right)/\sigma_{ij}\right\}$ and covariates $\mathbf{x}_{ij}$ being the ones selected for the modelling of T_{ij} in equation (6). The rest of transition probabilities are as in equation (1). In this case, the marginal $p_{22}(\cdot)$ with regard to the conditional distribution of T_{12} given $T_{12} \leq s$ is

$$\begin{aligned}
p_{22}(s,t \mid \mathbf{x},\boldsymbol{\theta}) = \frac{\sigma_{12}^{-1}\lambda_{12}(\mathbf{x}_{12})}{1-\exp\{-\lambda_{12}(\mathbf{x}_{12})s^{\sigma_{12}^{-1}}\}} \int_0^s \Big(&u^{\sigma_{12}^{-1}-1} \\
&\times \exp\left\{-\lambda_{12}(\mathbf{x}_{12})u^{\sigma_{12}^{-1}} - \lambda_{23}(\mathbf{x}_{23})\left((t-u)^{\sigma_{23}^{-1}} - (s-u)^{\sigma_{23}^{-1}}\right)\right\}\Big)\mathrm{d}u \quad (8)
\end{aligned}$$

Note that in our fitted model $\sigma_{12} = \sigma_{13}$ and equations in (7) simplify.

Given covariates, all these probabilities depend on the parameters of the disability model, which joint posterior distribution factorizes given in our case as

$$\pi(\boldsymbol{\theta} \mid \text{data}) = \pi(\boldsymbol{\theta}_{13}, \boldsymbol{\theta}_{12} \mid \text{data})\,\pi(\boldsymbol{\theta}_{23} \mid \text{data}) \quad (9)$$

where posteriors $\pi(\boldsymbol{\theta}_{12},\ \boldsymbol{\theta}_{13} \mid \text{data})$ and $\pi(\boldsymbol{\theta}_{23} \mid \text{data})$ have been approximated by *MCMC* as discussed in Section 4. Consequently, Bayesian reasoning is able to propagate all the information in equation (9) to obtain the posterior distribution $\pi(p_{ij}(s,\ t \mid \mathbf{x},\ \boldsymbol{\theta}) \mid \text{data})$ for transition probabilities. Note that we can also consider the posterior distribution for the transition probabilities given T_{12} by changing the marginal in equation (8) for the corresponding $p_{22}(s,\ t \mid \mathbf{x},\ \boldsymbol{\theta},\ t_{12})$ in equation (7).

The posterior distribution for transition probabilities is a very useful and important tool which provides direct information about not only the possible values of the transition probabilities but also their uncertainty. Approximation of each posterior $\pi(p_{ij}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) \mid \text{data})$ is obtained by plugging the *MCMC* draws $\{\boldsymbol{\theta}^{(m)}, m = 1, \ldots, M\}$ of the posterior distribution of $\boldsymbol{\theta}$ in equation (9) into conditional transition probabilities, where $p_{12}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ and $p_{22}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ are approximated, for each draw of $\boldsymbol{\theta}$, using adaptive quadrature integration. Moreover, we can approximate summaries of the above posterior distribution using Monte Carlo sums. For example, the expectation of the posterior distribution $\pi(p_{ij}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) \mid \text{data})$ can be approximated by

$$\mathrm{E}(p_{ij}(s, t \mid \mathbf{x}, \boldsymbol{\theta}) \mid \text{data}) \approx \frac{1}{M} \sum_{m=1}^{M} p_{ij}(s, t \mid \mathbf{x}, \boldsymbol{\theta}^{(m)})$$

Other posterior relevant summaries such as variances, medians and credible intervals for $p_{ij}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ can be approximated in a similar way.

Our inferential procedure is general and of course it allows the computation of the posterior transition probabilities for any patient with known specific values for the covariates. In fact, one of the aims of the project is to construct a software product to automatically compute relevant information about the possible evolution of a patient. Next, and in order to illustrate the summaries about the posterior distribution for the transition probabilities using the available data, we show some results for three generic patient profiles:

- Baseline-risk patient, where all continuous covariates are equal to the mean of the sample and the discretized covariates have been fixed to values corresponding to a low risk.
- High-risk patient, for which continuous covariates values are set to one SD above the mean for variables that increase the risk, as the calcaemia, and one below the mean for protective factors as the albumin. Similarly, discretized covariates have been fixed to the highest risk setup.
- Low-risk patient, where conditions are opposite, in a symmetric way, to the high-risk patient.

Table 5 shows a summary of the covariate values considered for the definition of the three different types of patients.

In order to summarize and appropriately visualize the relevant information of the posterior distribution for each transition probability, all of them are graphically presented in terms of their posterior mean and 95% credible interval.

For a baseline-, high- and low-risk patient, Figure 2 shows the first year of evolution from diagnosis (upper row) and from progression (lower row), i.e. conditioning on $T_{12} = s$. Note that the time scale in both lower-row graphics is not time since diagnosis but it is easy to recover it simply by adding the observed $T_{12} = s$ value to the current time since progression. Figure 3 refers to a period of a year starting after 6 months in Stage IV without progression (upper row) and after 6 months of

Table 5. Values of the covariates for each type of patient considered

Patient	Alb	Cal	Bmi	Cea	Ldh1	Ldh2
Baseline risk	3.41	9.67	0	0	1	0
High risk	2.72	10.08	1	1	0	1
Low risk	4.10	9.26	0	0	0	0

evolution assuming that the cancer has progressed at an unknown time $T_{12} \leq 180$ (bottom row), i.e. using the marginal probabilities given in equation (8), for baseline-, high- and low-risk patients. The arrangement of the plots inside each figure emulates the transition probability matrix. Note that, consequently, the posterior mean of the distributions in each row sum to one.

In general, in all types of patients, posterior means for $p_{11}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$, conditional $p_{22}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, T_{12} = s)$ and marginal $p_{22}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$ decrease with time while posterior expectation for $p_{13}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$, $p_{23}(s, t \mid \mathbf{x}, \boldsymbol{\theta}, T_{12} = s)$ and $p_{23}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$ increase. This is consistent with the general fact that the expected probability of dying increases with time but it is more pronounced for high-risk patients. Figure 2 shows that posterior means for $p_{12}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ rise in a very similar way for both baseline and low-risk patients while we expect a non-monotonically behavior in high-risk patients. For these patients, we observe a strong downfall in $p_{11}(0, t \mid \mathbf{x}, \boldsymbol{\theta})$ after around the 2 month of treatment, and, after practically the first 3 months of treatment, a high-risk patient is expected to either progress or die. The same general comments apply to posterior distributions for all transition probabilities from state 1 in Figure 3.

It can also be appreciated in both figures that the width of the 95% credibile intervals usually grows over time and that it is larger the larger t is. The variability of the posterior for transition probabilities from state 2 is slightly higher than the one corresponding to transitions from state 1,

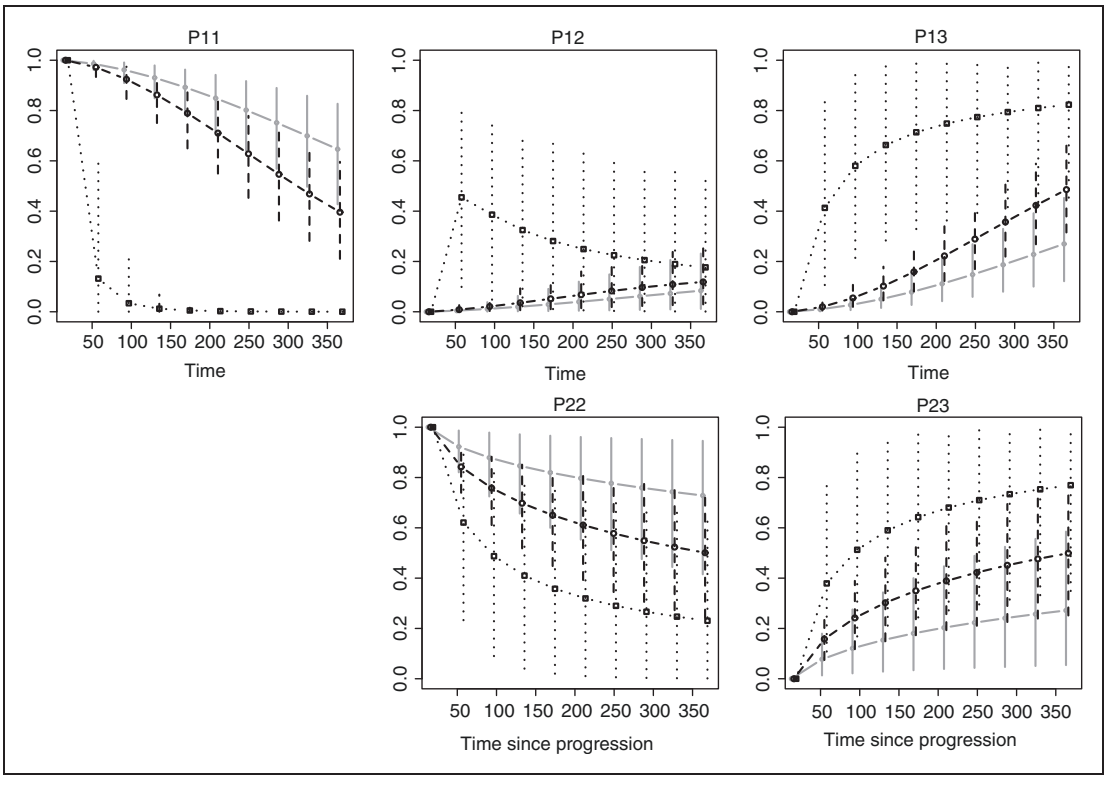

Figure 2. Posterior mean (points) and corresponding 95% credible interval (vertical segments) of all probability transitions for a baseline- (- - - ○ - - -), high- (··· □ ···) and low-risk (—●—) patients during the first year of treatment (upper row) and during a year after progression (bottom row).

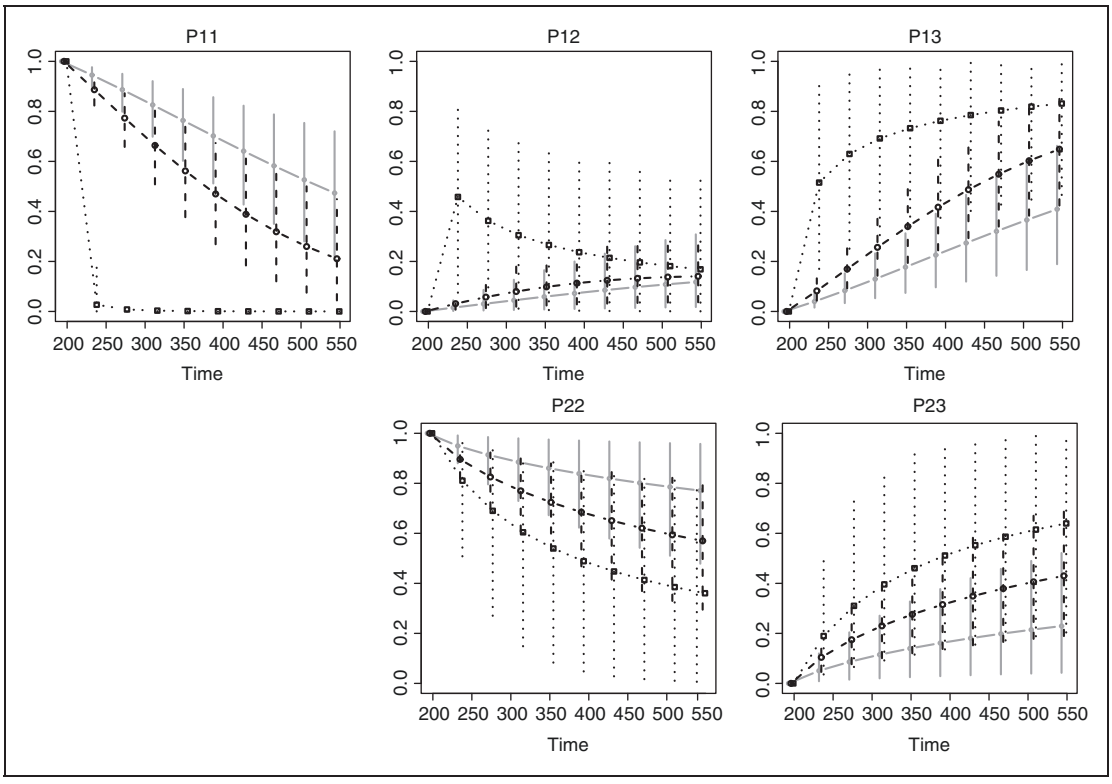

Figure 3. Posterior mean (points) and corresponding 95% credible interval (vertical segments) of all probability transitions for a baseline- (- - - ○ - - -), high- (··· □ ···) and low-risk (—•—) patient for a year after the first 6 months of treatment (upper row) and after 6 months of evolution assuming that the tumour has progressed at $T_{12} \leq 180$ ($p_{2j}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$, $j = 2, 3$) (bottom row).

due to the small number of patients that have progressed in the observed sample. Finally, variabilities of the posterior distributions for high-risk patients are higher with respect to baseline- and low-risk patients, also due to the small number of high-risk patients in the sample; for instance, only the 9% of the patients have a value for Cea greater than 30 (Table 2).

Comparing the posterior distribution for transition probabilities $p_{1j}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ when $s = 0$ (Figure 2) and $s = 180$ days (Figure 3), posterior means for $p_{11}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ are lower and for $p_{13}(s, t \mid \mathbf{x}, \boldsymbol{\theta})$ are higher with $s = 180$ days in average trend. On the other hand, we observe the opposite behaviour of the posterior distributions for conditional p_{2j} (Figure 2) and marginal p_{2j} (Figure 3), giving more death expectations to patients in the conditional situation.

For patients that progress in $T_{12} \leq 180$ (bottom row in Figure 3), the posterior distributions for $p_{22}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$ have in general a greater mean value than for $p_{23}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$. The only exception is for risk patients for whom we can appreciate that the posterior distributions for transition probability to death has a greater mean value than the posterior for remaining in the progression state when $t \geq 400$, although with a large uncertainty.

Regarding a baseline-risk patient starting at $s = 0$ (Figure 2), we can observe that posterior probabilities of moving from state 1 to state 2 are very low; in fact after the first year, these probabilities remain practically smaller than 0.2. When $s = 180$ days (Figure 3), transition $1 \rightarrow 3$

becomes more likely. In the case that the patient is in state 2 when $s = 180$, the probability of remaining in state 2 is above 0.5 for another 6 months. After that time, for around $t \geq 400$ posterior probability for $p_{22}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$ and $p_{23}(180, t \mid \mathbf{x}, \boldsymbol{\theta})$ are both compatible with 0.5.

For high-risk patients, when $s = 0$, Figure 2 indicates that the posterior probability mean associated to sojourn time in the initial state decreases dramatically in the first 50 days, being very low, almost 0, for $t \geq 100$. This results in an increase of the posterior mean for $p_{12}(0, t \mid \mathbf{x}, \boldsymbol{\theta})$ and $p_{13}(0, t \mid \mathbf{x}, \boldsymbol{\theta})$ of the same range. After $t = 50$ days, the posterior mean for $p_{12}(0, t \mid \mathbf{x}, \boldsymbol{\theta})$ begins to decrease to about 0.25 after a year.

Summarizing, for a high-risk patient in state 1, the probability of stay after 1 year is almost 0. However, it is not evident whether the patient is likely to progress or die, because both probabilities are compatible with 0.5 due to the large uncertainty. For the case of $s = 180$ in Figure 3, the posterior probability of remaining in state 1 goes practically to 0 in about 3 months, while the posterior mean for the transition probabilities from initial state to death are, in general, higher than that to progression, although with a large uncertainty.

Finally, in the case of low-risk patients, posterior mean for $p_{11}(0, 360 \mid \mathbf{x}, \boldsymbol{\theta})$ is around 0.7, while $p_{13}(0,360 \mid \mathbf{x}, \boldsymbol{\theta})$ is around 0.2 (Figure 2). For the case $s = 180$ (Figure 3), the sojourning in state 1 is the most probable event, followed by the transition to state 3. The posterior probabilities of sojourning in state 2, a year since progression (Figure 2), are also greater than the posterior probabilities of transition to state 3; the mean posterior probability for $p_{22}(s, s + 360 \mid \mathbf{x}, \boldsymbol{\theta}, T_{12} = s)$ is around 0.8 while the posterior mean for $p_{23}(s, s + 360 \mid \mathbf{x}, \boldsymbol{\theta}, T_{12} = s)$ is around 0.2.

6 Conclusions

This article presents the development of a multi-state semi-Markov regression model that incorporates measured baseline covariates to explain the variation in panel data. Inference is done from a Bayesian perspective which allows flexible modelling and the estimation of the distribution for transition probabilities. Such type of inference allows to properly account for the uncertainty in the data, because it relies on conditional inference.

In particular, we consider a disability model for assessing the evolution of Stage IV *NSCLC* patients with times between transitions modelled using Weibull regression analysis. The Weibull survival model has proved to be flexible enough to accommodate the shape of the oncologists' expected survival functions. Although non-parametric models are very popular in survival analysis, as stated in Ibrahim et al.,[32] parametric models play an important role in Bayesian survival analysis, since they offer straightforward inference, especially useful for small sample sizes. The methodology used for model selection combines the *BIC* together with the expert knowledge, leading to an effective way of extracting the prognostic factors for each of the survival models used. We acknowledge that the use of *BIC* having a small sample size can be problematic, and therefore, expert knowledge has to be considered with the aim to mitigate this drawback.

From a methodological point of view, the main contribution of our study is that it provides a basis for making individual estimations in the case where there is no prior information available or when it is not trustable via the use of minimum informative priors. Of course, more data and more accurate information about the follow-up of those patients could enrich the model, allowing for the possibility of including time covariates, interval censoring or treatment informations, obtaining thus more accurate results and providing a solid ground for the assessment of model adequacy and validation.

Within a medical perspective, it is worth emphasizing that, in addition to the widely accepted prognostic factors, Cea, Bmi, Alb and Ldh were also found to be significant prognostic factors for

time to progression.[33,34] An interesting and previously unexplored finding is that the prognosis is particularly poor for Stage IV *NSCLC* patients with values of calcaemia varying in the normal range. This may reflect either a greater burden of tumour cells within the bone marrow, or the effect of a yet not described chemokine or cytokine secreted by the tumour into the circulation. It is known that hypercalcaemia is a poor prognostic factor in patients with lung cancer.[35,36] However, little has been studied about association between serum corrected calcium in the normal range and survival in advanced lung cancer patients.

The usefulness of this study for clinical research and clinical practice is compelling in this population of Stage IV *NSCLC* patients, in which advances in treatment have been slow with only modest improvements in survival. Data regarding Cea, Bmi, Alb, Ldh and Cal are readily available for all patients, which makes our model a clinically feasible metric.

Acknowledgements

The authors thank the two anonymous referees for their helpful comments that improved the original version of this article.

Funding

This study has been partially supported by the Ministerio de Ciencia e Innovación [grant number MTM2010-19528], Mutua Madrileña [grant AP75942010], Ministero dell'Istruzione, dell'Universitá e della Ricerca of Italy and the visiting professor program of the Regione Autonoma della Sardegna.

References

1. Sant M, Allemani C, Santaquilani M, et al. EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary. *Eur J Cancer* 2009; **45**(6): 931–991.
2. Langer CJ, Besse B, Gualberto A, et al. The evolving role of histology in the management of advanced non-small-cell lung cancer. *J Clin Oncol* 2010; **28**: 5311–5320.
3. Brown BW, Brauner C and Minnotte MC. Noncancer deaths in white adult cancer patients. *J Natl Cancer Inst* 1993; **85**(12): 979–987.
4. Andersen PK, Borgan P, Gill RD, et al. *Statistical models based on counting processes*. New York: Springer, 1993.
5. Keiding N, Klein JP and Horowitz MM. Multi-state models and outcome prediction in bone marrow transplantation. *Stat Med* 2001; **20**: 1871–1885.
6. Andersen PK and Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2002; **11**: 91–115.
7. Foucher Y, Giraland M, Soulillou JP, et al. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Stat Med* 2007; **26**(6): 5381–5393.
8. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, et al. Multi-states models for the analysis of time to event data. *Stat Methods Med Res* 2009; **18**: 195–222.
9. Porta N, Calle ML, Malats N, et al. A dynamic model for the risk of bladder cancer progression. *Stat Med* 2012; **31**: 287–300.
10. Van Den Hout A and Matthews FE. Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model. *Stat Med* 2008; **27**(26): 5440–5455.
11. Siannis F, Farewell VT and Head J. A multi-state model for joint modelling of terminal and non-terminal events with application to Whitehall II. *Stat Med* 2007; **26**(2): 426–442.
12. Kneib T and Hennerfeind A. Bayesian semiparametric multi-state models. *Stat Model* 2008; **8**: 169–198.
13. Van Den Hout A and Matthews FE. Estimating dementia-free life expectancy for Parkinson's patients using Bayesian inference and microsimulation. *Biostatistics* 2009; **10**: 729–743.
14. Hougaard P. Multi-state models: a review. *Lifetime Data Anal* 1999; **5**: 239–264.
15. Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007; **26**: 142–161.
16. Klein JP and Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. New York: Springer, 2003.
17. Berger J. The case for objective Bayesian analysis. *Bayesian Anal* 2006; **1**(3): 385–402.
18. Volinsky CR and Raftery AE. Bayesian information criterion for censored survival models. *Biometrics* 2000; **56**(1): 256–262.
19. Rossi A, Maione P, Bareschino MA, et al. The emerging role of histology in the choice of first-line treatment of advanced non-small cell lung cancer: implication in the clinical decision making. *Curr Med Chem* 2010; **17**: 1030–1028.
20. Soon YY, Stockler MR, Askie LM, et al. Duration of chemotherapy for advanced non-small-cell lung cancer: a systematic review and meta analysis of randomized trials. *J Clin Oncol* 2009; **27**: 3277–3283.
21. Jeremic B, Milicic B, Dagovic A, et al. Pretreatment clinical prognostic factors in patients with stage IV non-small cell lung cancer (NSCLC) treated with

chemotherapy. *J Cancer Res Clin Oncol* 2004; **129**: 114–122.
22. Oakes D. Multiple time scales in survival analysis. *Lifetime Data Anal* 1995; **1**(1): 7–18.
23. Pencina MJ, Larson MG and D'Agostino RB. Choice of time scale and its effects on significance of predictors in longitudinal studies. *Stat Med* 2007; **26**: 1343–1359.
24. Andersen PK and Pohar Perme M. Inference for outcome probabilities in multi-state models. *Lifetime Data Anal* 2008; **14**(4): 405–431.
25. Lee PN and O'Neill JA. The effect both of time and dose applied on tumor incidence rate in benzopyrene skin painting experiments. *Br J Cancer* 1971; **25**: 759–770.
26. Doll R. The age distribution of cancer: implications for models of carcinogens. *J R Stat Soc, Ser A* 1971; **134**: 133–166.
27. Evans IG and Nigm AM. Bayesian prediction for two-parameter Weibull lifetime models. *Comm Stat Theor Meth* 1980; **9**(6): 649–658.
28. Albert J. *Bayesian computation with R*. New York: Springer, 2009.
29. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
30. Kass RE and Raftery AE. Bayes factors. *J Am Stat Assoc* 1995; **90**: 773–795.
31. Gilks WR, Richardson S and Spiegelhalter DJ. *Markov chain Monte Carlo in practice*. London: Chapman & Hall, 1996.
32. Ibrahim JG, Chen M and Sinha D. *Bayesian survival analysis*. New York: Springer, 2001.
33. Win T, Sharples L, Groves AM, et al. Predicting survival in potentially curable lung cancer patients. *Lung* 2008; **186**: 97–102.
34. Forrest LM, McMillan DC, McArdle CS, et al. A prospective longitudinal study of performance status, an inflammation based score (GPS) and survival in patients with inoperable non-small cell lung cancer. *Br J Cancer* 2005; **92**: 1834–1836.
35. Stewart AF. Hypercalcemia associated with cancer. *N Engl J Med* 2005; **352**(4): 373–379.
36. Hiraki A, Ueoka H, Takata I, et al. Hypercalcemia-leukocytosis syndrome associated with lung cancer. *Lung Cancer* 2004; **43**(3): 301–307.

Appendix

The full likelihood of the semi-Markov illness–death model is constructed through the counting process representation of the likelihood for a multi-state process given in Andersen and Keiding.[6] In order to facilitate the comprehension of the results, we have also omitted here all the conditioning parameters and covariates for the hazard rate, density and survival functions for the relevant transition times.

We consider different type of patients and express the contribution of each of them to the full likelihood:

- Type 1 patient

Patient k is diagnosed at $y^I_{1,k} = 0$ in Stage IV, cancer does not progress and is still living at the end of the follow-up period, r_k, thus being $y^O_{1,k} = r_k$ and $c_{12,k} = c_{13,k} = 0$.

$$\begin{aligned} L(\text{patient } k) &= \exp\left(-\int_0^{r_k} h_{12}(u)\,\mathrm{d}u\right)\exp\left(-\int_0^{r_k} h_{13}(u)\,\mathrm{d}u\right) \\ &= S_{12}(r_k)\,S_{13}(r_k) \\ &= f_{12}(y^O_{1,k})^{c_{12k}} S_{12}(y^O_{1,k})^{1-c_{12k}} f_{13}(y^O_{1,k})^{c_{13k}} S_{13}(y^O_{1,k})^{1-c_{13k}} \end{aligned}$$

where recall that $h_{ij}(\cdot)$, and $f_{ij}(\cdot)$ are, respectively, the hazard rate function and the probability density function for T_{ij}, transition time between states i and j, and

$$S_{ij}(u) = \exp\left(-\int_0^u h_{ij}(u)\,\mathrm{d}u\right)$$

- Type 2 patient

Patient k is diagnosed at $y^I_{1,k} = 0$ in Stage IV, cancer does not progress but he/she dies during the follow-up at $t_{13,k}$. Then, $y^O_{1,k} = y^I_{3,k} = t_{13,k}$, and $c_{13,k} = 1$ and $c_{12,k} = 0$ because the only observed transition is $1 \to 3$.

$$
\begin{aligned}
L(\text{patient } k) &= \exp\left(-\int_0^{t_{13k}} h_{12}(u)\,\mathrm{d}u\right) h_{13}(t_{13,k}) \exp\left(-\int_0^{t_{13k}} h_{13}(u)\,\mathrm{d}u\right) \\
&= S_{12}(t_{13,k}) f_{13}(t_{13,k}) \\
&= f_{12}(y_{1,k}^O)^{c_{12k}} S_{12}(y_{1,k}^O)^{1-c_{12k}} f_{13}(y_{1,k}^O)^{c_{13k}} S_{13}(y_{1,k}^O)^{1-c_{13k}}
\end{aligned}
$$

- Type 3 patient

Patient k is diagnosed at $y_{1,k}^I = 0$ in Stage IV, the cancer progresses at $t_{12,k}$ but she/he is still living at the end of the follow-up period $r_k = t_{12,k} + (r_k - t_{12,k})$. As transition $1 \to 2$ occurs in $t_{12,k}$, $y_{1,k}^O = y_{2,k}^I = t_{12,k}$, $c_{12,k} = 1$ and $c_{13,k} = 0$. Since the patient remains alive at the final of the follow-up, $(r_k - t_{12,k})$ will be the duration of time that patient k has been seen at state 2, from $t_{12,k}$ to r_k. So, $y_{2,k}^O = r_k$ and $c_{23,k} = 0$.

$$
\begin{aligned}
L(\text{patient } k) &= h_{12}(t_{12,k}) \exp\left(-\int_0^{t_{12k}} h_{12}(u)\,\mathrm{d}u\right) \exp\left(-\int_0^{t_{12k}} h_{13}(u)\,\mathrm{d}u\right) \\
&\times \exp\left(-\int_{t_{12k}}^{r_k} h_{23}(u - t_{12,k})\,\mathrm{d}u\right) \\
&= f_{12}(t_{12,k})\, S_{13}(t_{12,k})\, S_{23}(r_k - t_{12,k}) \\
&= f_{12}(y_{1,k}^O)^{c_{12k}}\, S_{12}(y_{1,k}^O)^{1-c_{12k}} f_{13}(y_{1,k}^O)^{c_{13k}}\, S_{13}(y_{1,k}^O)^{1-c_{13k}} \\
&\times f_{23}(y_{2,k}^O - y_{2,k}^I)^{c_{23k}}\, S_{23}(y_{2,k}^O - y_{2,k}^I)^{1-c_{23k}}
\end{aligned}
$$

- Type 4 patient

Patient k is diagnosed at $y_{1,k}^I = 0$ in Stage IV, cancer progresses at $t_{12,k}$ and finally dies during the follow-up at $t_{12,k} + t_{23,k}$. Transition to progression at $t_{12,k}$ implies that $y_{1,k}^O = y_{2,k}^I = t_{12,k}$, $c_{12,k} = 1$ and $c_{13,k} = 0$. Also, transition from progression to death at $t_{12,k} + t_{23,k}$ results in $y_{2,k}^O = t_{12,k} + t_{23,k}$ and $c_{23,k} = 1$.

$$
\begin{aligned}
L(\text{patient } k) &= h_{12}(t_{12,k}) \exp\left(-\int_0^{t_{12k}} h_{12}(u)\,\mathrm{d}u\right) \exp\left(-\int_0^{t_{12k}} h_{13}(u)\,\mathrm{d}u\right) \\
&\times h_{23}(t_{23,k}) \exp\left(-\int_{t_{12k}}^{t_{12k}+t_{23,k}} h_{23}(u - t_{12,k})\,\mathrm{d}u\right) \\
&= f_{12}(t_{12,k})\, S_{13}(t_{12,k}) f_{23}(t_{23,k}) \\
&= f_{12}(y_{1,k}^O)^{c_{12k}}\, S_{12}(y_{1,k}^O)^{1-c_{12k}} f_{13}(y_{1,k}^O)^{c_{13k}}\, S_{13}(y_{1,k}^O)^{1-c_{13k}} \\
&\times f_{23}(y_{2,k}^O - y_{2,k}^I)^{c_{23k}}\, S_{23}(y_{2,k}^O - y_{2,k}^I)^{1-c_{23k}}
\end{aligned}
$$

The full likelihood is the product of the likelihood for all patients, thus obtaining the expression (2), where the set $\mathcal{P}$ is defined as the subset of indexes corresponding to type 3 and type 4 patients.